

December 2009

Enhancing User Access to Microdata in Australia

Access to microdata assists and encourages informed decision making through enabling wider use of ABS data for social and economic research and analysis. The ABS has been making microdata available, under certain conditions, for statistical purposes in the form of Confidentialised Unit Record Files (CURFs) since 1985. The ABS has a well-established program for producing and providing access to CURFs and continues to develop strategies to improve researcher access to, and use of, unit record data for such purposes. The ABS currently releases three types of microdata: Basic CURFs, Expanded CURFs and Specialist CURFs, via three modes of access: CD-ROMs, the Remote Access Data Laboratory (RADL) and the ABS on-site Data Laboratory (ABSDL).

Today there exists high user demand for the ABS to provide access to more detailed unit record data in a more flexible way, across a wider array of datasets (such as business data and longitudinal linked datasets). An inability to meet these demands will increasingly become a disadvantage to ABS core business, the relevance of the ABS and ultimately to the coherence of the NSS. This, along with a number of other drivers for change including the growing risk of identification, has led the ABS to propose a new strategy for accessing ABS microdata into the future.

The proposed future strategy is to:

- continue to produce and release basic CURFs for use in the users' environment;
- progressively replace RADL with a remote execution environment for microdata (REEM), primarily for table generation and basic statistical analysis; and
- to increase the use of the ABSDL for complex analysis of microdata, including providing access to longitudinal and linked datasets.

The key components of the REEM are the development of a Survey Table Builder (similar to the Census TableBuilder) and an Analysis Service. It is proposed that these services will access detailed de-identified microdata, with confidentiality routines built into the outputs generated to ensure that they are confidentialised in line with ABS legislative requirements and can be released as public use outputs (that is, they can be published and shared with others without restrictions).

REEM will use internationally recognised standards for the exchange of data and metadata including the use of the Data Documentation Initiative (DDI) and Statistical Data and Metadata Exchange (SDMX) and machine-to-machine interfaces (APIs), enhancing the ABS's leadership role in national and international statistics. The use of DDI/SDMX will also serve as a catalyst for the ABS strategic intent to use DDI/SDMX for internal systems.

The first stage of REEM, development of a Survey Table Builder for Household collections, has commenced with staged development planned to occur over the next few years. In parallel, research has commenced into exploring future directions for the analysis service. Through December 2009 the ABS will be undertaking a re-assessment of user requirements and consulting with other producers of microdata in Australia.

If you would like any further information on the proposed changes, or would like to be consulted during the assessment phase please contact Melissa Gare on (02) 6252 7147 or m.gare@abs.gov.au.

Cost Modelling for Economic Surveys

The Operations Research Unit (ORU) is undertaking a number of projects into modelling the cost of intensive follow-up (IFU) processes for ABS economic surveys. These projects are part of a conceptual framework which links input parameters and processes, such as survey sample size, number of reminder letters to be sent, number of calls to be made to providers, to cost and statistical outcomes, such as bias of output estimates and response rates. Currently we are developing statistical models to estimate these inputs more robustly, and to link these inputs to outcomes, to enable predictions to be made, and to inform decision making on what operational practice efficiencies can be made.

The first model which we are developing is a probability of response model. The aim of this model is to determine the drivers of response patterns and to be able to predict response rates resulting from various IFU process scenarios, leading to an improved understanding of IFU best practice. For example, the model will be able to be used to inform the optimal number of calls to make, and when to make them, and thus will link into other problems such as optimal interviewer allocation, and minimising the dollar cost of operations for an

acceptable level of statistical quality of the output estimates.

This model will make use of survey paradata for a number of surveys and cycles. Survey paradata includes information on individual businesses' response status over the survey cycle, how many reminder letters and subsequent follow-up calls were made to providers, how many incoming calls were received from providers, and what the relative timing of each of these were, as determined by the IFU contact strategy. It also includes business characteristics such as business industry and size and their priority of contact within the IFU strategy.

Because the timing of events is an important part of the IFU process, survival analysis is being used to fit the probability of response model. Survival analysis is a technique commonly used to model time-to-event data, where some records may be censored (i.e. the timing of the event is unknown or is outside the observation window) and where the explanatory variables may vary over time. In our case, the event is receiving a response from the business (non-respondents are censored), and some of the explanatory variables (e.g. number of calls) vary over the analysis period. Such a model will allow us to predict not only the end of enumeration response rate, but also the response rate throughout the enumeration period.

The other models in the conceptual framework are the Provider Contact Unit (PCU) cost model and a soon to be developed bias model. The PCU cost model, which has been reviewed and enhanced by the ORU, can be used to forecast costs based on survey characteristics (such as sample size), a specified IFU strategy, and expected response rates over time. For the bias model, as proof of concept we are using a simulated bias based on the practical application of reducing the response rate. In the near future, a more robust approach to modelling the bias will be investigated.

Once the three component models have been linked together, PCU will be able to make better informed decisions on practical issues, for example, reducing target response rates, or how long to wait between follow-up calls. Early next year, ORU will commence work on the application of scientific methods to determine optimal practices and procedures, such as interviewer allocation/ scheduling and the number of calls which should be made to a provider.

For more information, please contact Melanie Black on (02) 6252 7241 or melanie.black@abs.gov.au

Predicting Travel Costs in Household Surveys

In addition to modelling the cost of intensive follow-up processes for ABS economic surveys (see related article above), the Operations Research Unit (ORU) is also undertaking a number of projects into modelling the cost of interview processes for ABS household surveys. These projects are also part of a conceptual framework which links input parameters and processes, such as survey sample size, interviewer characteristics, and

follow-up effort to cost and statistical outcomes, such as bias of output estimates and response rates. Currently we are developing statistical models to estimate these inputs more robustly, and to link these inputs to outcomes, to enable predictions to be made, and to inform decision making on what operational practice efficiencies can be made.

As part of this work, the ORU has been developing a model to predict the cost of interview travel time and motor vehicle allowance (MVA) for household surveys. Initially this is being done for the Monthly Population Survey (MPS). Currently these costs account for about 40% of the total enumeration cost in standard workloads.

This analysis makes use of MPS paradata. Interviewers record individual visits to each household, with the outcome of that visit (e.g. interview, noncontact, etc) recorded on the CAIWMS (Computer Assisted Interviewer Workload Management System). These call-level records can be combined to identify interviewer call patterns at a broader level, such as the number of times an interviewer visits each household, and the number of times they visit a particular block. Matching these with records of the actual time and distance claimed for a trip provides a dataset that can be used to develop models of time and distance as a function of interviewer call patterns. Factors found to be good predictors of time and distance were calls made to households, trips made to blocks, and number of fully-responding households.

Various factors could cause bias in the fitted model. In particular, because of quality limitations of the paradata, not all calls can be matched with time/distance records, causing missingness that may lead to bias. Bias can be reduced by calibrating cost predictions for the fitting months against the actual survey costs recorded by PSO for those months to produce calibration factors that can then be used to correct predictions for future months.

To validate this time/distance model, it was fitted and calibrated on MPS data from July 2007 to April 2008, and then tested on call records from June-December 2008, months that included the MPS sample cut. Predictions were actually better in the validation months than for the fitting months. This is due to a large number of missing/unmatched records for August and September 2007 that caused errors in prediction for these months. The actual MVA costs between June and December 2008 decline with the sample cut, and the predicted model costs reliably track this pattern, staying within about 7% of the cost at each time point and matching movements from month to month.

Besides use for budget forecasting, this model can also be used to dissect costs and determine how different activities affect MPS costs and therefore influence the choice of efficient survey operations.

For example, some MPS households require visits after the first month, either because the interviewer wasn't able to establish telephone contact or because the occupant has requested face-to-face interviews. Applying this model showed that although these non-first-month-in-sample (NFMIS) households produce only 35% of face-to-face interviews, these

households account for approximately 54% of travel/MVA costs. This is because first-month households are clustered, allowing an interviewer to work efficiently by visiting several neighbouring households in one trip. A NFMIS household is likely to be the only one in its block that needs to be visited face-to-face, requiring the interviewer to make a special trip just for that one household.

One practical example of informed decision making on efficient processes would be the cost of increasing the response rates by allowing further interview contacts. By comparing cost and response at each call, it's possible to estimate the marginal cost per interview at different parts of the survey, giving the time and expenditure required to secure one more respondent. As the survey goes on, this cost increases because the chances of response become poorer, and because it's less efficient to visit a small number of remaining non-respondents. For example, in the first month in sample, making the first visit to each MPS household costs about \$28 for each response achieved but by the fifth call, the ABS is spending approximately \$60 for each additional response. Under a limited budget, it is never possible to secure 100% response rate, but comparing marginal costs in different parts of the sample helps show ways to achieve the largest possible response within that budget.

More recent work includes an investigation of the effects of geography on interviewer travel time and distance, in combination with factors discussed above. ORU determined that the distance between the workload and the interviewer's home is an important predictor of travel time and distance in most cases. However, when the workload is more than 100km from the interviewer's home, the exact distance becomes unimportant, because the interviewer is generally working from a temporary base rather than travelling from home.

MPS costs and statistical quality are both closely linked to response rates. Therefore, ORU is currently working on a standardised way to link these phenomena to allow for "big picture" decision-making in household surveys. In the future, ORU aims to extend the cost model and linked model to other household surveys in the SSS program.

For further information, please contact Geoffrey Brent on (03) 9615 7685 or geoffrey.brent@abs.gov.au.

Completion of Backcasting Effort for Classification-Impacted Subannual Time Series

The introduction of several revised classification schemes, including the ANZSIC 2006 (A06) industry classification, into ABS subannual collections this year impacted on a large number of time series. As these are measurement changes and are not related to any real world changes, classification changes to the characteristics of our time series needed to be managed so that we can continue to publish consistent time series which are comparable across time and more useful for analysis and seasonal adjustment purposes.

To achieve this, base A06 time series were created from data collected under the old ANZSIC 1993 (A93) scheme and calculated via macro concordance tables, in such a way as to be as close as possible to the new A06 classification. For many collections, parallel samples were then collected under both A93 and A06-based sample designs, and used to estimate the impact of A06 implementation. To bring the estimates from the two schemes into agreement, this impact was then smoothed out across the length of each time series via the Generalised Backcasting Facility (GBF). The GBF performs these backcasts in such a way as to minimise the revisions to seasonally adjusted movement estimates, while preserving relevant aggregative relationships between time series.

The Time Series Analysis section (TSA) was involved in the planning of A06 implementation and the creation of the GBF, and performed the backcasts for most collections. For areas which performed their own backcasts, TSA offered advice and undertook work to help assure the quality of the backcast time series. As of December 2009, these backcasts and quality assurance tasks have been completed.

From the release of September quarter data in November and December 2009, all relevant ABS subannual collections will be published solely on an A06 basis. As there is currently insufficient data to detect changes in the seasonal behaviour of time series due to A06 implementation, TSA will continue to monitor the series, particularly as it returns to its regular schedule of annual seasonal reanalysis (ASR) of collections from 2010 onwards.

For more information, please contact Chris Mann on (02) 6252 6758 or chris.mann@abs.gov.au.

Treatment of Non-Response in ABS Business Surveys

Item and unit non-response is a problem that affects most ABS business surveys, leading to an increase in the sampling error and bias of the estimates. The two standard methods for the treatment of non-response are imputation and weighting, with the former being more commonly used for item non-response and the latter for unit non-response.

Many ABS business surveys have traditionally used a weight adjustment method equivalent to using a "live respondent mean" impute for the non-respondent units. This particular weight adjustment was implemented into ABS business surveys prior to the availability of useful auxiliary information and the introduction into estimation systems of the generalised regression (GREG) estimator, which makes efficient use of this auxiliary information. While this non-response treatment was considered the best possible solution in the situation where there was no auxiliary information available, it is inefficient if useful auxiliary information exists which can be used to reduce the sampling error and non-response bias.

An alternative two-phase calibration approach to non-response has been proposed, where the sampling

mechanism is considered the first phase and the response mechanism is considered the second phase. The proposed two-phase calibration estimator was based on auxiliary information being available for all units in the population and the live/death status of a unit being known for units selected in the survey. It has been shown that if useful auxiliary information exists, then the generalised regression estimator leads to considerable sample size reductions, and the two-phase calibration approach to non-response leads to further sample size reductions.

An evaluation on the Annual Integrated Collection (AIC) demonstrated that the proposed two-phase calibration estimator can achieve reductions in the relative standard errors in the order of 12.5% and potentially reduce the non-response biases from 1.2% to 0.3% for the key data items. Alternatively, the proposed two-phase calibration estimator could be used to achieve considerable sample size savings of approximately 1,000 units, while at the same time achieve reductions in the relative standard errors in the order of 4%.

It is a relatively simple exercise to implement the proposed two-phase calibration estimator into the estimation systems used for ABS business surveys, as well as implement modifications to the existing rescaled bootstrap procedure, so that it is possible to produce approximate variances for the two-phase calibration estimator. The current estimation systems consist of a set of modular components that perform specific processes which can be easily interconnected to produce estimates from a range of estimators. The development of several simple weighting components will enable the estimation systems to produce estimates and variances for the proposed two-phase calibration estimator.

A paper on the proposed two-phase calibration estimator, recently presented to the Methodology Advisory Committee, will soon be available on the ABS website.

For more information, please contact John Preston on (07) 3222 6229 or john.preston@abs.gov.au.

Children's Participation in Sports

Participation in organised cultural and physical activities is an important element of a child's social development. In recent years, increasing awareness of the incidence of childhood obesity has particularly highlighted the desirability, on health grounds, for children to participate in regular physical activity. Participation in organised sport, as a subset of broader physical activity, is also important for the development of motor coordination skills, teamwork and physical fitness. While changes in children's sports participation over their life cycle can be best examined using longitudinal data, which follows the same children over time, such data are not always available and are often costly to collect. One alternative is to pool data from successive cross-sectional surveys and construct a 'pseudo-longitudinal' dataset to study this phenomenon.

The Analytical Services Branch recently released a research paper entitled 'Children's Participation in

Organised Sporting Activity' that looked at children's participation in organised sports within a 'pseudo-longitudinal' framework (ABS Cat. No. 1351.0.55.028). The study was based on data from three repeated cross-sectional surveys of Children's Participation in Cultural and Leisure Activities (2000, 2003 and 2006). It examined how participation in sports by children changes as they grow older and what factors influence such participation, including the effects of age, period and cohort. While these three independent cross-sectional surveys do not allow tracking of participation of individual children over time, the aligning of the age, period and cohort characteristics allow the study to follow the average behaviour of children from the same birth-year cohort, and compare their age-specific participation rates with those of other cohorts, while controlling for changing social and cultural influences.

The authors initially applied a simple 'age-period-cohort' accounting model to the full dataset and to selected subpopulations of interest to get some insights into the relative importance of these dimensions in explaining the variation in participation rates. Evidence of strong age effects was found, indicating that children's participation in organised sporting activity reaches a peak in the 9–11 years age-group, and then declines. This behaviour was consistently observed, even for groups of children that report significantly different rates of participation. A rising trend in participation rates between 2000 and 2006 was also observed, but this was not uniformly observed over all subpopulations. In particular, no increase in participation was reported among children from more disadvantaged areas. No discernible evidence of consistent cohort effects, however, was detected.

To obtain further insights into the factors influencing participation in organised sporting activity, especially at the individual level, the authors then fitted a logistic regression model to the data, supplementing the age, period and cohort effects with a range of observed socio-demographic characteristics pertaining to the child, the family and the neighbourhood. Not only did this modelling framework allow for the identification of the relative importance of key factors associated with children's participation in sport, it also provided a means of verifying the preliminary findings on age, period and cohort effects. The results from the earlier analysis suggested an appropriate solution to the identification problem which arises when age, period and cohort variables are included simultaneously in a model. The cohort variable was collapsed into fewer categories, focussing attention on the stronger age and period effects.

Results from the logistic regression model confirmed the significant age and period effects, and factors such as gender, parents' employment status, country of birth and the relative socioeconomic status of the neighbourhood were found to be strongly associated with children's participation rates. Children who spent more time on television and computers were also found to be less likely to participate in organised sporting activities. No cohort effects in sports participation were observed during the period under study, confirming the preliminary findings.

The absence of a cohort effect in the model was not totally unexpected – given the short period spanned by the study and the minimal age separation of the defined cohorts. An initial objective of the study was to assess whether or not younger cohorts of children are participating less in sport than their predecessors. The limitations of this study, however, would suggest that it is premature to form a strong view on this question. The main shortcoming of the current analysis is that most cohorts have not yet been tracked over all age-groups, and hence it is not possible to contrast their age-specific participation rates. The inclusion of data from future surveys (e.g. the 2009 CPCLA) should lead to improvements in the estimation of cohort effects. There remain, of course, questions that can only be answered by true longitudinal data. For example, it is not possible to observe whether young children who played sport will continue to play sport, or what factors may prompt individual children to engage or disengage in a range of sporting or alternative activities.

For more information, please contact Anil Kumar on (02) 6252 5344 or anil.kumar@abs.gov.au.

Adjusting Longitudinal Datasets: Analysis of the Simulated SLCD

The Statistical Longitudinal Census Dataset (SLCD) aims to link a 5% sample of Census 2006 records to their 2011 Census counterparts using Census fields with the notable omission of names and addresses. In preparation, a Simulated SLCD was conducted, linking 2005 Census Dress Rehearsal (CDR) records to 2006 Census records.

As some records inevitably contain missing or incorrect fields, and some fields legitimately change over time, a proportion of records will be linked incorrectly or will remain unlinked. The simulated SLCD offered a chance to assess the impact of errors in linking, and to explore adjustment methods which may be employed to minimise their impact on analysis. As it was conducted during the Census processing period, names and addresses were available for use in creating a linked Simulated SLCD. This additional linking information was used to create a high quality Gold Standard (GS) linked dataset. The Bronze Standard (BS), which was created without names and addresses, better resembles the linking process planned for the SLCD, and was evaluated against the GS. The GS is set to be deleted in December 2009, now that this evaluation has been completed.

The BS was compared against the GS, but additionally the adjustment techniques were applied to the BS and evaluated. The BS was weighted using two methods. First by benchmarking to the CDR; and second by modelling the probability a record linked on the GS is linked on the BS and using the inverse of these probabilities as weights. Weighting aimed to correct primarily for bias introduced by records which were unlinked on the BS. Additionally, adjustment techniques were developed, aiming to correct for both incorrect links and unlinked records. These methods were based on maximum likelihood techniques, and

again used the GS to estimate the probability of being linked or linked correctly.

Evaluating both the unadjusted BS and adjustment approaches consisted of running analyses to answer example research questions of interest. Results suggest that the BS yields the same conclusions as the GS for most analyses. However, the BS was inadequate for analysis of people living in indigenous communities, as it performed poorly in linking these individuals. Weighting tended not to improve the BS, however moderate (and in some cases large) improvements were gained using the newly developed adjustment techniques.

Looking forward, plans are underway for a second simulated SLCD, linking the 2010 CDR with the 2011 Census, before the SLCD is conducted. Properties of the simulated SLCD BS need to be carefully considered when drawing inferences about the SLCD. The linking methodology will differ in the SLCD, and linking files five years apart will produce challenges not present in linking files one year apart. An information paper describing weighting is currently being written, which will describe the options for weighting the SLCD.

For more information, please contact Paul Campbell on (02) 6252 7101 or paul.campbell@abs.gov.au, or James Chipperfield on (02) 6252 7301 or james.chipperfield@abs.gov.au.

How to Contact Us and Subscriber Emailing List

The Methodological Newsletter features articles and developments in relation to work done within the ABS Methodology and Data Management Division. By its nature, the work of the Division brings it into contact with virtually every other area of the ABS. Because of this, the newsletter is a way of letting all areas of the ABS know of some of the issues we are working on and help information flow. We hope the Methodological Newsletter is useful and we welcome comments.

If you would like to be placed on our electronic mailing list, please contact:

Jayne McQualter
Methodology & Data Management Division
Australian Bureau of Statistics
Locked Bag No. 10
BELCONNEN ACT 2617

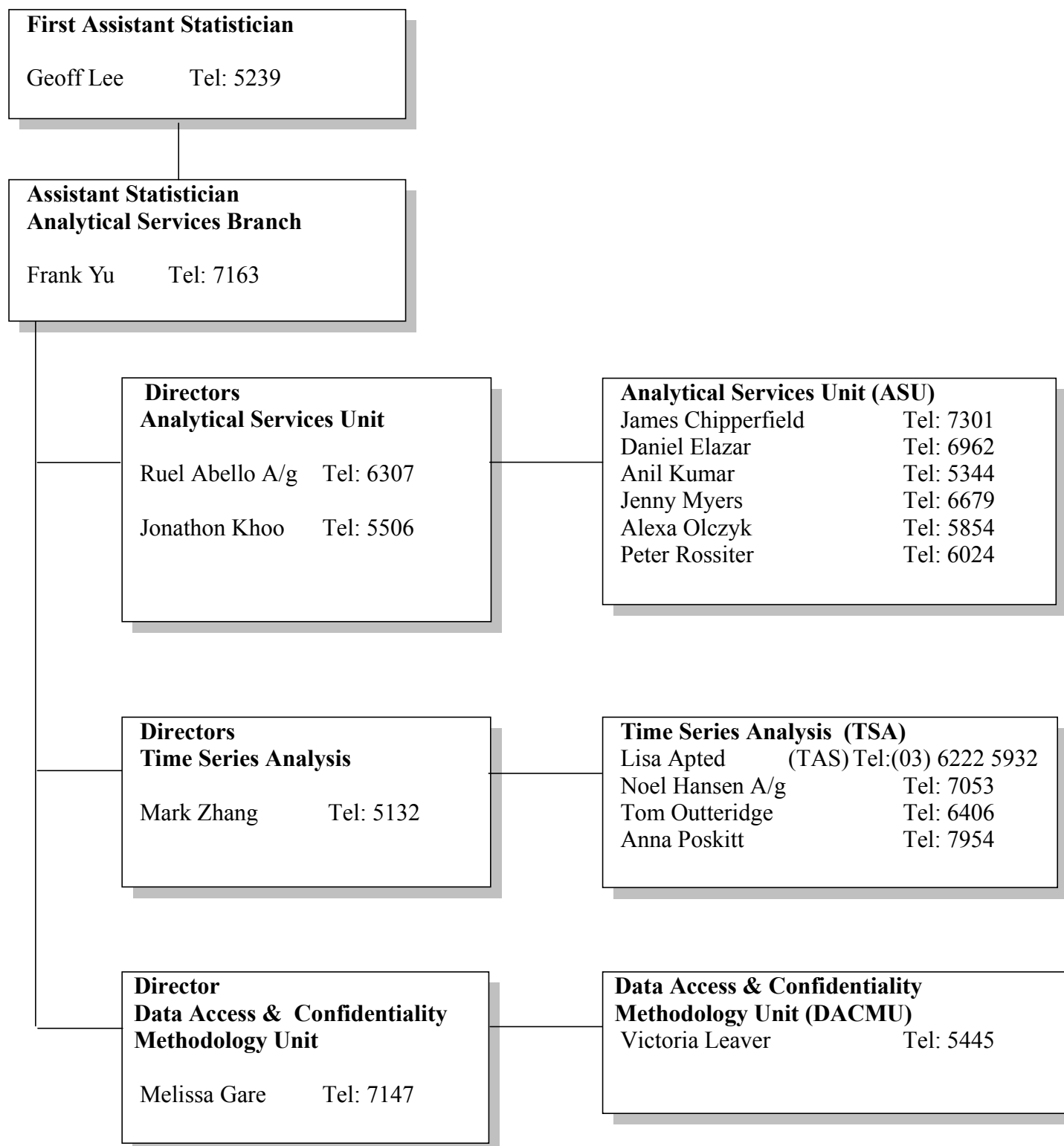
Tel: (02) 6252 7320
Email: methodology@abs.gov.au

Click on the following links to view the ABS Privacy Statement and Disclaimer
[Privacy Statement](#) | [Disclaimer](#)

Methodology & Data Management Division

Management Structure

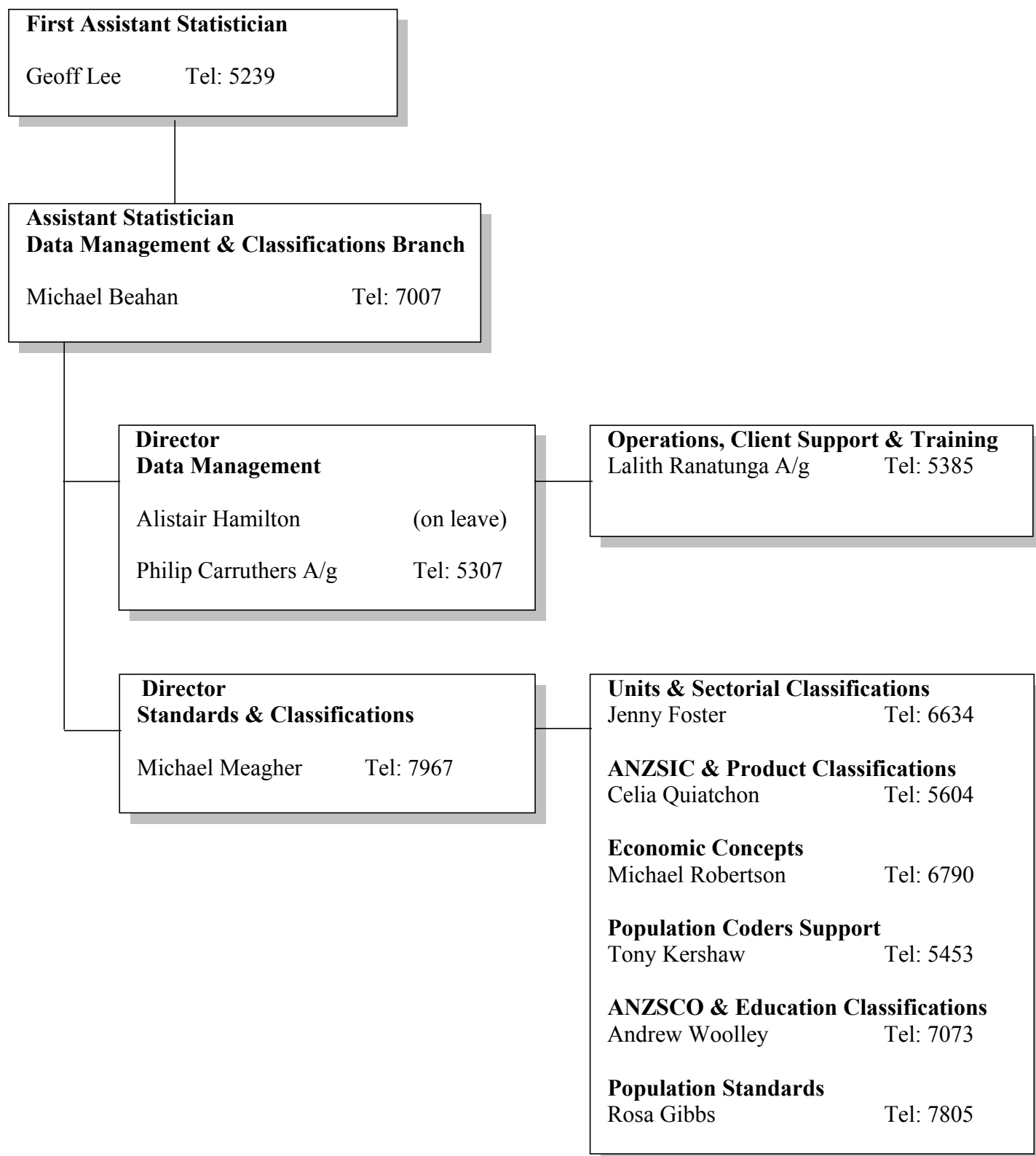
Current at December 2009



Methodology & Data Management Division

Management Structure

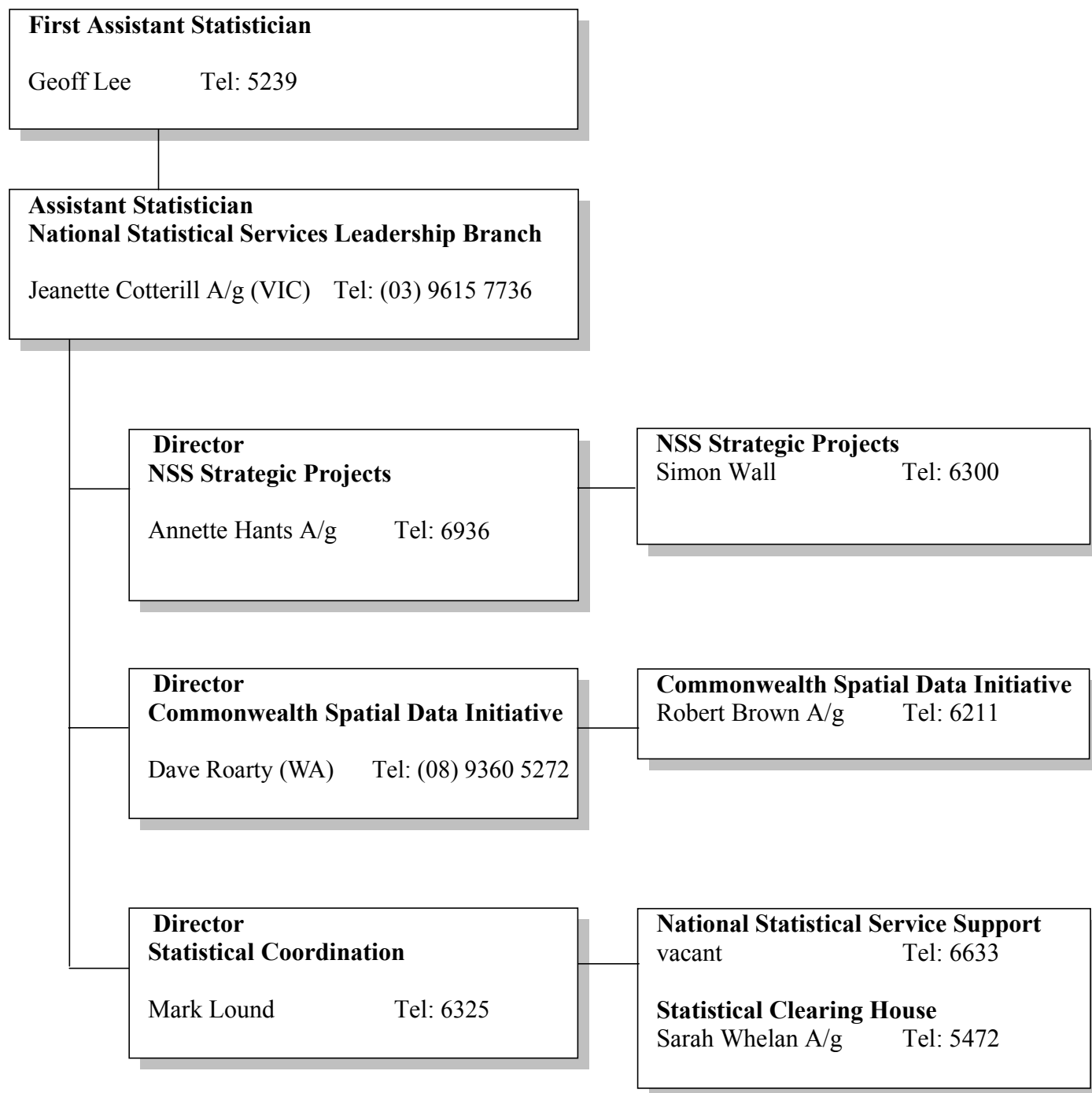
Current at December 2009



Methodology & Data Management Division

Management Structure

Current at December 2009



Methodology & Data Management Division

Management Structure

Current at December 2009

